Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 166–169, January 1987
Evolution

The guanine and cytosine content of genomic DNA and bacterial evolution

(biased mutation pressure/codon usage/neutral theory)

AKIRA MUTO AND SYOZO OSAWA

Laboratory of Molecular Genetics, Department of Biology, Faculty of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464, Japan

Communicated by Motoo Kimura, September 15, 1986

ABSTRACT **The genomic guanine and cytosine (G+C) content of eubacteria is related to their phylogeny. The G+C content of various parts of the genome (protein genes, stable RNA genes, and spacers) reveals a positive linear correlation with the G+C content of their genomic DNA. However, the plotted correlation slopes differ among various parts of the genome or among the first, second, and third positions of the codons depending on their functional importance. Facts suggest that biased mutation pressure, called A·T/G·C pressure, has affected whole DNA during evolution so as to determine the genomic G+C content in a given bacterium. The role of A·T/G·C pressure in diversification of bacterial DNA sequences and codon usage patterns is discussed in the perspective of the neutral theory of molecular evolution.**

Among eubacteria, the mean guanine and cytosine (G+C) content of genomic DNA varies from approximately 25% to 75%. That the bacterial genomic G+C content is somehow related to phylogeny has been suggested (1, 2). The phylogenetic tree of eubacterial 5S rRNA clearly indicates this relationship (3). According to this tree Gram-negative and Gram-positive bacteria separated earliest. Among Gram-positive bacteria, those with low genomic G+C content such as *Bacillus subtilis* (genomic G+C, 42%), *Lactobacillus viridescens* (40%), *Staphylococcus aureus* (33%), *Clostridium perfringens* (38%), and *Mycoplasma capricolum* (25%) are phylogenetically close, whereas those with high genomic G+C, such as *Micrococcus luteus* (75%), *Streptomyces griseus* (73%), and *Mycobacterium tuberculosis* (67%), comprise one phylogenetic group. These two groups separated long ago. Gram-negative bacteria with intermediate G+C content, such as *Escherichia coli* (50%), *Serratia marcescens* (58%), *Salmonella typhimurium* (51%), and *Pseudomonas fluorescens* (60%) belong to the common Gram-negative branch.

These facts suggest that the differences in genomic G+C content may have been caused by mutation pressure—the direction and magnitude of this pressure varying among the phylogenetic lines. Such mutation pressure, which we call biased A·T/G·C pressure due to biased mutation rates among the four bases, seems to have been exerted on the entire genome during evolution. Mutations caused by A·T/G·C pressure are subject to selective constraints that usually operate in the form of negative selection to eliminate functionally deleterious changes. A certain fraction of non-deleterious (i.e., neutral) mutations are then fixed in the population due to random genetic drift (4). Thus, functionally less important parts in the genome evolve faster than more important ones in accordance with the neutral theory of molecular evolution (see ref. 5). In other words, for a given species, the A·T/G·C pressure changes the G+C content of various parts of the genome in the same direction, but to different extents depending on their functional importance.

A large amount of data on the DNA sequences of different genes and spacers from various bacteria are now available. Using these data, we present here lines of evidence that support the above view. The study further suggests that A·T/G·C pressure has played a major role in diversification of genomic DNA sequences and codon usage in bacterial evolution.

Correlation of G+C Content Between Entire Genome and the Specific Parts of Genome

The bacterial genome is roughly composed of protein genes (70–80%), spacers including various signals (20–30%), and stable RNA genes (<1%). Fig. 1 shows a plot of the G+C content of different parts of the genome vs. the mean G+C content of various bacterial genomes and indicates that all components positively but differentially correlate to the genomic G+C content. In 1962 Miura (6) demonstrated that among bacteria the G+C content of rRNA (16S + 23S + 5S rRNAs) is about the same, whereas that of tRNA weakly correlates to genomic G+C content. There exists a weak positive correlation of the G+C content of both rRNA and tRNA to genomic G+C content. However, the G+C content of spacers and protein genes reveals a strong linear correlation with genomic G+C content: among various bacteria the G+C content of spacers ranges from about 20% to 80% and that of protein genes ranges from about 30% to 75%, as genomic DNA G+C content varies from 25% (*M. capricolum*) to 75% (*M. luteus*). Thus, for a given bacterium, the G+C content of spacers, protein genes, and stable RNA genes is all biased in the same direction as the G+C content of the total genome.

This bias is stronger for spacers, less so for protein genes, and least for stable RNA genes, although their contribution to the average G+C content is in the order of protein genes, spacers, and stable RNA genes. These positive colinealities shown in Fig. 1 strongly support the idea that the A·T/G·C pressure strongly influenced the G+C content of all DNA during evolution. The differential levels of G+C content in the different components of the genome in a given organism can be understood as the consequence of selective constraints that have been exerted on the genome to eliminate functionally deleterious mutants. Since most parts of spacers are functionally the least important in the genome, the larger part of mutations in these regions is selectively neutral, and therefore the evolutionary rate is higher than for other parts. The rRNA and tRNA genes are less variable, because their transcripts are nontranslatable, and most, if not all, of their sequences are important for biological functions. Protein genes are more variable than the stable RNA genes, because larger parts of synonymous codon changes and conservative amino acid changes occur without deleterious effects.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

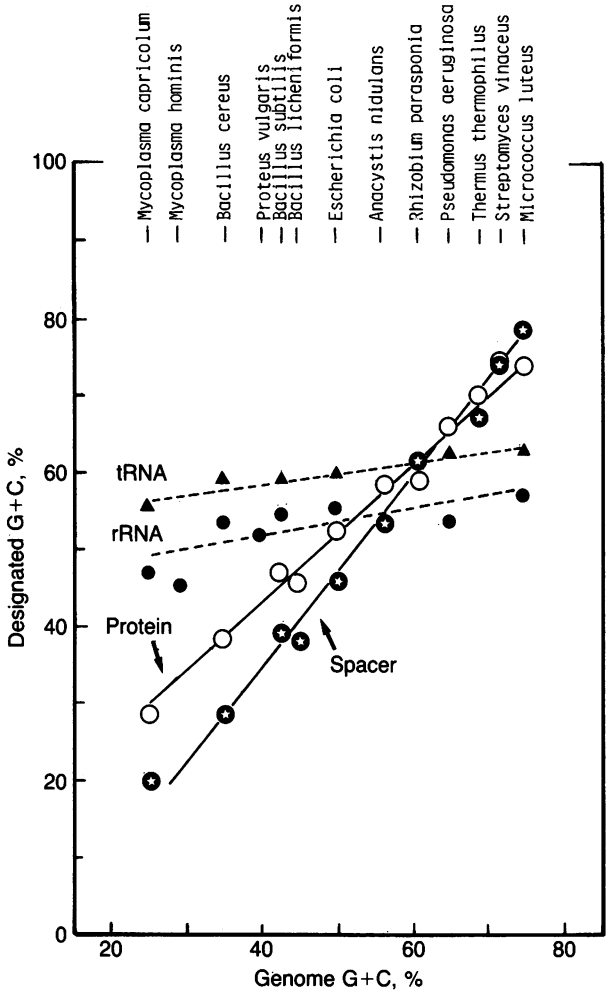

FIG. 1. Correlation of G+C content between total genomic DNA and designated parts of the genome. Values for rRNA and tRNA are taken from Miura (6), Midgley (7), and Maniloff and Morowitz (8). References to protein genes are as follows: ribosomal protein genes of *M. capricolum* (9); penicillinase gene of *B. cereus* (10); 21 genes of *B. subtilis* collected by Ogasawara (11); penicillinase gene of *B. licheniformis* (12); 27 genes of *E. coli* collected by Koningsberg and Godson (13); ribulose-1,5-biphosphate carboxylase/oxygenase gene of *A. nidulans* (14); nitrogenase gene of *R. parasponia* (15); mercuric reductase gene of *P. aeruginosa* (16); isopropylmalate dehydrogenase gene of *T. thermophilus* (17); viomycin phosphotransferase gene of *S. vinaceus* (18); ribosomal protein-encoding genes of *M. luteus* (T. Ohama, F. Yamao, A. M., and S.O., unpublished data). Values of spacers are calculated from the 5′- and 3′-flanking sequences of the above genes.

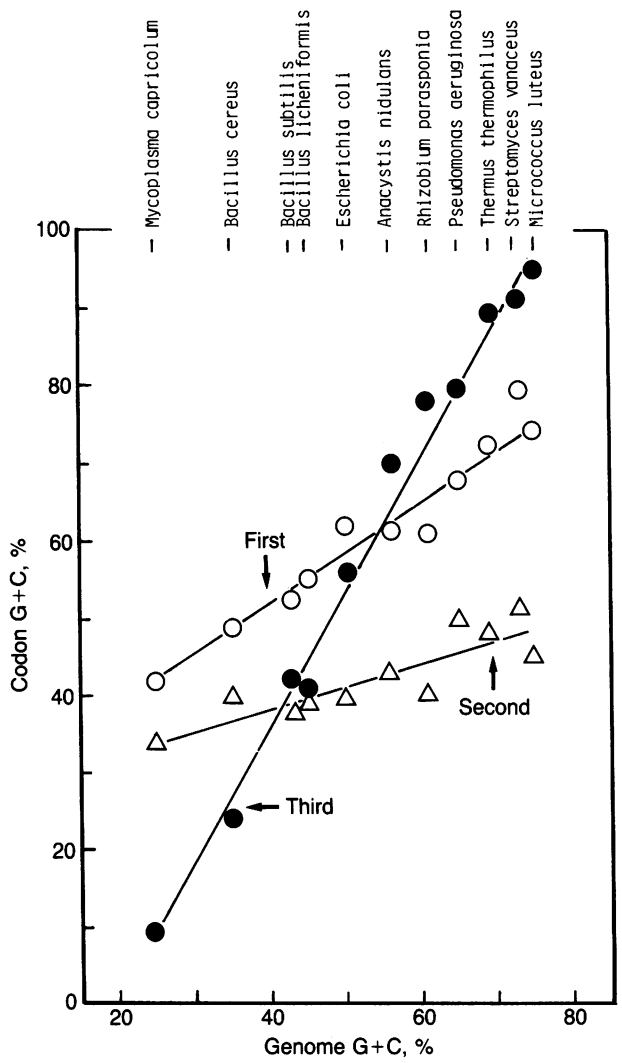

FIG. 2. Correlation of the G+C content between total genomic DNA and the first, second, and third codon positions. Values are taken from the references in the legend for Fig. 1.

Codon Usage

A nonrandom feature of codon usage has been widely recognized (19, 20). In Fig. 2 are plotted the G+C contents of the first, second, and third codon positions, respectively, from various bacterial species vs. G+C content of the corresponding genome. The G+C contents of all three positions reveal a linear positive relationship with genome G+C content, although slopes differ—the steepness rank order being third, first, and second codon positions. Most striking is that the G+C content of the third positions varies linearly from about 10% in *M. capricolum* to >90% in *M. luteus*. The strong correlation is thus due to A·T/G·C substitutions at the codon third position.

In the present study, comparisons are made between different protein genes. It is preferable to use homologous genes for this purpose, but this kind of available data is limited at present. The most valuable genes for such comparisons may be those for ribosomal proteins. Functional structure of ribosomes is well conserved throughout the eubacterial group, and hence most, if not all, of the codon- and amino acid-substitutions would be predicted to be neutral or nearly neutral. In other words, most of the observable substitutions would have no selective advantages for the ribosomal functions. Thus, we compared substitutions at 1188 homologous codon sites in the genes for nine ribosomal proteins in *spc* operon between *M. capricolum* (genome G+C, 25%; ref. 21 and unpublished work) and *E. coli* (50%; ref. 22). Table 1 summarizes the codon substitution patterns. Between the two species, 199 codons are identical. There exist 360 synonymous (silent) codon substitutions, 169 conservative amino acid substitutions (see legend for Table 1), and 459 other substitutions between them. About 81% of the synonymous codon substitutions occurs at the third position. In most cases, codons ending in guanine or cytosine in *E. coli* are replaced by a synonymous codon ending in adenine or uracil(thymine) in *M. capricolum* (see also ref. 21). About 17% of the synonymous changes occurs at the first plus third positions, and all of them result in an adenine- or uracil(thymine)-richness in *M. capricolum* as compared with *E. coli*. Among 169 conservative amino acid substitutions, which occur primarily by changing the codon first and/or third positions, 114 codon changes occur in the direction to higher A+U(T)-content in *M. capricolum*. For example,

Table 1. A+T* substitutions at 1188 homologous codon sites of ribosomal protein-encoding genes† between *M. capricolum* and *E. coli*

Codon position	Substituted codons, no.												
	Synonymous‡				Conservative§				Others				
	Gain	Loss	No gain/loss	Subtotal	Gain	Loss	No gain/loss	Subtotal	Gain	Loss	No gain/loss	Subtotal	Total
1	1	0	0	1	16	1	6	23	22	9	12	43	67
2	0	0	0	0	0	3	9	11	7	6	0	13	24
3	201	14	77	292	4	0	6	10	9	1	7	17	319
1 + 2	0	0	5	5	1	0	0	1	29	11	20	60	66
1 + 3	61	0	0	61	53	1	28	82	45	11	17	73	216
2 + 3	0	0	0	0	13	0	1	14	43	10	26	79	93
1 + 2 + 3	1	0	0	1	27	2	0	29	110	35	29	174	204
Total	264	14	82	360	114	6	50	170	265	83	111	459	989 (1188)¶
No. of changed positions	327	14	87	428	235	11	79	325	602	185	232	1019	1772
Net A + T gain‖	+303	−14	0	+289	+177	−6	0	+165	+401	−91	0	+310	+764

Of the 1188 codon sites, 199 are identical.
*Number of codons that gain or lose A+T in *M. capricolum* as compared with *E. coli*.
†Nine ribosomal protein-encoding genes (S5, S8, S14, L5, L6, L14, L15, L18, and L24) in *spc* operon of *E. coli* (22) and *M. capricolum* (ref. 21; unpublished data).
‡Synonymous (silent) codon substitution.
§Conservative amino acid substitution includes Lys/Arg, Leu/Ile, Leu/Val, Ile/Val, Ser/Thr, Ala/Gly, Gln/Asn, Glu/Asp and Phe/Tyr.
¶Inclusive of identical codons.
‖Net A+T gained in *M. capricolum* as compared with *E. coli*.

many CGC or CGU codons for arginine, or CUG for leucine in *E. coli* are replaced by AAA for lysine or AUU for isoleucine in *M. capricolum*. Among 459 other codon substitutions, which are also probably neutral for the reason previously discussed, a gain of A+U(T) is observed in 256 codons of *M. capricolum*, whereas a loss of A+U(T) is observed in 83 codons. Among 989 substituted codons out of the total 1188 codons, 643 codons (65%) gain A+U(T), 103 (10%) lose A+U(T), and 243 (25%) show neither gain nor loss of A+U(T). The total of A+U(T) gain and loss in all nucleotide sites (2967) in the substituted codons is 875 and 111, respectively.

All these facts indicate that a strong biased mutation pressure has caused *M. capricolum* to discriminate against guanine and cytosine and to use adenine and uracil wherever possible. In contrast to the A+T-biased codon usage in *M. capricolum*, a strongly G+C-biased codon usage has been shown in G+C-rich bacteria, such as *Streptomyces* species (18, 23) and *M. luteus* (unpublished data). Thus, we conclude that this biased mutation pressure is the major dictator for codon usage in these bacteria. Ikemura (19, 20) demonstrated a positive correlation between codon usage and tRNA abundance in *E. coli* and in yeast, claiming that tRNA population is the major factor providing selective constraints on codon use. Kagawa *et al.* (17) stressed that the high G+C content in the codon third positions in an extremely thermophylic bacterium, *Thermus thermophilus*, is the consequence of high temperature selection, because G·C pairs are more thermostable than A·T pairs in the double-stranded structure of DNA. This explanation cannot be applied to other G+C-rich bacteria, such as *Micrococcus, Streptomyces*, etc., which are not thermophylic. Thus, the results shown in Fig. 2 and Fig. 1 strongly suggest that biased A·T/G·C pressure predominates over other selective forces in determining nonrandom codon usage.

Discussion

More than twenty years ago, Sueoka (24) presented a theory to account for diversification of the G+C content of the bacterial genome; the G+C content of DNA of a given bacterium will be determined by the effective base conversion rate u (G·C to A·T) and v (A·T to G·C). The G+C content of equilibrium (p) is $v/(v + u)$. When u/v is 3.0, 1.0, and 0.33, the G+C content of the bacterium will be 25% (e.g., *M. capricolum*), 50% (*E. coli*), and 75% (*M. luteus*), respectively. In this theory only the mutation pressure in two directions—G·C to A·T and A·T to G·C—is assumed. Thus, our A·T/G·C pressure is equivalent to Sueoka's u/v. Remember, however, that at each nucleotide site, one of the four bases is fixed most of the time and that such a concept of equilibrium can be applied only when a very large number of sites are considered. Superimposed on mutation pressure, mutational changes are subject to selective constraints, and a certain fraction of nondeleterious mutations can become fixed in a given population (4, 5).

The nature of biased A·T/G·C pressure is unknown. One possibility is that mutational modification of some components in DNA synthesis biases the mutation rate towards either A·T pairs (25) or G·C pairs (26). Another possibility is that methylation and deamination of DNA bases lead to the accumulation of A·T pairs in DNA (27, 28); the abundance of methylation/deamination enzymes and/or the activity of the deamination–repair system could thus be candidates for biasing mutation pressure.

Biased mutation pressure acts on the entire genome. This is best deduced from the genome structures of the extremely G+C-poor bacterium *M. capricolum*, in which all parts of the genome (protein genes, stable RNA genes, and spacers) contribute systematically but to a different extent to the low G+C content of the genome (21, 29–31). This is also the case for G+C-rich bacteria such as *M. luteus* (unpublished data) and *Streptomyces* species (18, 23), in which all parts of the genomes contribute to high genomic G+C content. These facts, together with Figs. 1 and 2, strongly suggest that biased mutation pressure has been exerted uniformly on all of the DNA. Different evolution rates at different parts of a given genome can best be explained by the negative selection principle of neutral theory (4, 5, 32, 33)—i.e., functionally less important parts of the genome have evolved more rapidly than more important ones. The most striking example is the fast evolutionary rate of spacer regions (Fig. 1) and the synonymous codon substitution that occurred predominantly at codon third positions (Fig. 2). The correlation of mean G+C content of genomic DNA to bacterial phylogeny also

favors the view that A·T/G·C pressure has been exerted on the genome in bacterial evolution.

We have found that one of the chain termination codons, UGA (opal nonsense codon), in the universal code is read as tryptophan in *M. capricolum* (9, 31) and suggested that this change in the genetic code dictionary had occurred through biased mutation pressure. Jukes (34) has proposed a plausible pathway whereby UGA tryptophan codon evolved in *M. capricolum* without deleterious or lethal effects under a strong mutation pressure replacing G·C pairs by A·T pairs. He has further suggested that such biased mutation pressure could cause the genetic code to change in other organisms having biased DNA base composition. In fact, certain ciliates, in which a deviation from the universal genetic code has been discovered (35–39), have genomes with extremely low (≈30%) G+C content. The evolution of the genetic code in reference to A·T/G·C pressure requires further discussion (unpublished work).

We thank Dr. Motoo Kimura and our colleagues for valuable discussions. This work was supported by Grant-in-Aid from the Ministry of Education, Science and Culture of Japan.

1. Lee, K. Y., Wahl, R. & Barbu, E. (1956) *Ann. Inst. Pasteur (Paris)* **91,** 212–214.
2. Sueoka, N. (1961) *J. Mol. Biol.* **3,** 31–40.
3. Hori, H. & Osawa, S. (1986) *Biosystems* **19,** 162–172.
4. Kimura, M. (1968) *Nature (London)* **217,** 624–626.
5. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, UK).
6. Miura, K. (1962) *Biochim. Biophys. Acta* **55,** 62–70.
7. Midgley, J. E. M. (1962) *Biochim. Biophys. Acta* **61,** 513–525.
8. Maniloff, J. & Morowitz, H. J. (1972) *Bacteriol. Rev.* **36,** 263–290.
9. Muto, A., Yamao, F., Kawauchi, Y. & Osawa, S. (1985) *Proc. Jpn. Acad. B* **61,** 12–15.
10. Sloma, A. & Gross, M. (1983) *Nucleic Acids Res.* **11,** 4997–5004.
11. Ogasawara, N. (1985) *Gene* **40,** 145–150.
12. Neugebauer, K., Sprengel, R. & Schaller, H. (1981) *Nucleic Acids Res.* **9,** 2577–2588.
13. Koningsberg, W. & Godson, G. N. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 687–691.
14. Shinozaki, K., Yamada, C., Takahata, N. & Sugiura, M. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 4050–4054.
15. Scott, K. F., Rolfe, B. G. & Shine, J. (1983) *DNA* **2,** 141–148.
16. Brown, N. L., Ford, S. J., Primore, R. D. & Fritzinger, D. C. (1983) *Biochemistry* **22,** 4089–4095.
17. Kagawa, Y., Nojima, H., Nukiwa, N., Ishizuka, M., Nakajima, T., Yasuhara, T., Tanaka, T. & Oshima, T. (1984) *J. Biol. Chem.* **259,** 2956–2960.
18. Bibb, M. J., Bibb, B. J., Ward, J. M. & Cohen, S. N. (1985) *Mol. Gen. Genet.* **199,** 26–36.
19. Ikemura, T. (1981) *J. Mol. Biol.* **151,** 389–409.
20. Ikemura, T. (1982) *J. Mol. Biol.* **158,** 573–597.
21. Muto, A., Kawauchi, Y., Yamao, F. & Osawa, S. (1984) *Nucleic Acids Res.* **12,** 8209–8217.
22. Cerretti, D. P., Dean, D., Davis, G. R., Bedwell, D. M. & Nomura, M. (1983) *Nucleic Acids Res.* **11,** 2599–2616.
23. Gray, G. S. & Thompson, C. J. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 5190–5194.
24. Sueoka, N. (1962) *Proc. Natl. Acad. Sci. USA* **49,** 582–592.
25. Speyer, J. F. (1965) *Biochem. Biophys. Res. Commun.* **21,** 6–10.
26. Cox, J. F. & Yanofsky, C. (1967) *Proc. Natl. Acad. Sci. USA* **58,** 1895–1902.
27. Bird, A. P. (1980) *Nucleic Acids Res.* **8,** 1499–1504.
28. Dujon, B. (1981) in *The Molecular Biology of the Yeast Saccharomyces*, eds. Strathern, J. N., Jones, E. W. & Broach, J. R. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 505–635.
29. Iwami, M., Muto, A., Yamao, F. & Osawa, S. (1984) *Mol. Gen. Genet.* **196,** 317–322.
30. Sawada, M., Muto, A., Iwami, M., Yamao, F. & Osawa, S. (1984) *Mol. Gen. Genet.* **196,** 311–316.
31. Yamao, F., Muto, A., Kawauchi, Y., Iwami, M., Iwagami, S., Azumi, Y. & Osawa, S. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 2306–2309.
32. King, J. L. & Jukes, T. H. (1969) *Science* **164,** 788–798.
33. Kimura, M. & Ohta, T. (1974) *Proc. Natl. Acad. Sci. USA* **71,** 2848–2852.
34. Jukes, T. H. (1985) *J. Mol. Evol.* **22,** 361–362.
35. Caron, F. & Meyer, E. (1985) *Nature (London)* **314,** 185–188.
36. Preer, J. R., Jr., Preer, L. B., Rudman, B. M. & Barnett, A. J. (1985) *Nature (London)* **134,** 188–190.
37. Helftenbein, E. (1985) *Nucleic Acids Res.* **13,** 415–433.
38. Horowitz, S. & Gorovsky, M. A. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 2452–2455.
39. Kuchino, Y., Hanyu, N., Tashiro, F. & Nishimura, S. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 4758–4762.